Linguistic markers of secrets and sensitive self-disclosure in Twitter.

David J. Houghton
Birmingham Business School
University of Birmingham, UK.
d.j.houghton@bham.ac.uk

Adam N. Joinson
School of Management
University of Bath, UK.
A.Joinson@bath.ac.uk

Abstract

The present research sought to identify linguistic markers of sensitive self-disclosure in Twitter for three main purposes: (1) to support the development of software tools that can identify text as sensitive disclosure or not; (2) to contribute to the literature by establishing what is considered more sensitive disclosure in a specific CMC environment, and (3) to contribute to the methodological toolkit for studying sensitive self-disclosure. Two corpora were used in the present research. In Study 1 short messages were collected from Twitter and the site 'Secret Tweet' for comparison. In Study 2 'tweets' were collected and rated on sensitivity by six raters. LIWC and regression analyses were used to identify the linguistic markers of secret tweets (Study 1, 16 markers found) and sensitive self-disclosure (Study 2, 10 markers found). A software tool is developed to illustrate the markers in application. Implications for self-disclosure research, users, design and researchers are discussed.

Acknowledgements

This research was funded by an EPSRC grant to the PVNets project. (Grant#:EP/G002606/1). Thanks to Mathieu Roesch for his programming skills.

1. Introduction

The present research aims to identify the linguistic markers of sensitive self-disclosure in a specific form of computer-mediated communication (CMC) – Twitter - for three main purposes. First, to enable the development of software tools for use in social media that can identify user created text as sensitive or otherwise. Algorithms can then be applied to aid the user in their decision to share information with the different social spheres they are connected to, or to group relationships based on the sensitivity of information shared. Second, the identification of linguistic markers of sensitive self-disclosure will contribute to the self-disclosure literature by demonstrating what is considered more sensitive in a specific CMC context, and whether such markers are scalable to other media. Third, identifying linguistic markers will contribute to the development of methodological tools available to researchers studying interpersonal communication and privacy.

To address the first aim, we identify several practical methods to address the issue of over- or under- disclosure to different audience members. Software tools can be used systematically to identify sensitive self-disclosure in text. Procedures can then be developed to alert a user about the potential disclosure of sensitive information. However, this may benefit only those users who are unsure what is sensitive, e.g. younger users. The identification of sensitive text could be used alongside friend 'grouping' policies, such as those offered by Facebook and Google+, to ensure the desired group(s) of contacts receive sensitive posts. By default, new social media (NSM) sites could use these markers to automatically ensure the protection of users as opposed to presuming that all network contacts are the same. Users may be content with all their contacts receiving their status updates but can choose for more sensitive updates to be identified automatically and remove specified recipients.

This paper is outlined as follows. First, we outline the definition of self-disclosure and discuss the nature of 'sensitivity'. Second, we identify why self-disclosure is important in understanding use of NSM. We then address the measurement of self-disclosure and the need for an automated solution. Through two studies we identify the linguistic markers of sensitive disclosure in Twitter (and 'Secret Tweet'). Last, we demonstrate the use of the linguistic markers in a proof of concept application.

1.1. Self-Disclosure and Sensitivity

Self-disclosure can be defined as "that which occurs when A knowingly communicates to B information about A which is not generally known and

is not otherwise available to B" [1]. It can involve the sharing of knowledge between pairs of individuals, individuals within groups, between groups, or between individuals or groups and organizations [2]. Self-disclosure can be beneficial, having been linked to improved physical and psychological well-being, and increased trust and group identity [2, 3], and plays an important function in social relations – by reducing uncertainty [4], and increasing trust and liking in relationships [5, 6]. Self-disclosures are usually made for a purpose [7], including expression, self-clarification, social validation, relationship development and social control [7]. Individuals may fulfill these goals in different contexts and environments. The perceived appropriateness of a self-disclosure relates both to personal expectancies of individual behavior and social norms of behavior [4, 5, 7].

Self-disclosure can also be detrimental. For example, if the receiver of a disclosure believes the discloser intended to obtain social control rather than express themself, they may be interpreted as manipulative [7]. Much self-disclosure reveals vulnerability about the subject, and thus can become detrimental through the spread of information once it has been released or the misuse of that information by the receiver. In a sample of interviews, participants expressed concern that information disclosed in confidence had later been disseminated to others [8]. Other negative consequences can occur when too much or too little information is shared. Social exchange theory states that information is disclosed between individuals with increasing intimacy as conversation develops [5]. Sharing intimate information early in the conversation (& relationship) may result in the discloser being labeled a social deviant, with receivers suspicious as to their intentions [4, 5]. Similarly, it has been suggested that successful romantic relationships contain a degree of secrecy [9]. However, sharing too little information may result in conversation not reaching a higher level of intimacy and self-disclosure depth, hindering the development of a relationship, or signaling its dissolution [5].

Self-disclosure varies along two dimensions and 'too much disclosure' can relate to either breadth or depth of information [10]. Breadth relates to the quantity of information disclosure (both frequency and category), and depth to the quality [5, 10]. Depth can range from surface information e.g. biographic data, to deeper aspects including core beliefs and values [2, 5]. When considering the information people disclose about themselves, we should think of the person as comprising many different 'layers', akin to an 'onion' [5]. The central layers contain fewer aspects of the self (less breadth), but these aspects are deeper and more central to our view of ourselves [5]. Breadth can vary along two planes: frequency and category. Categories are the types of information disclosed, and frequency is the number of occurrences of disclosure [5].

In the present research we focus on the depth of self-disclosure. To develop linguistic markers that identify sensitive self-disclosure we are concerned less with how often (frequency) or with the strict category of message content, but more with the ability to identify deeper, more sensitive disclosure. While social exchange theory and social penetration theory [5] have received support in the self-disclosure literature e.g. [1, 11], it's unclear if the same core values are shared between new **social** media and other contexts, e.g. the typical core value of sexuality is one of the items often posted in the 'about me' section of NSM sites. In NSM salient core values may relate to the user's ability to manage social relations and their ability to create and maintain social capital. Therefore we expect that markers of sensitive information disclosure may relate to social relations rather than intrapsychic aspects of the self.

1.2. Self-Disclosure in New Social Media

NSM sites offer great potential for users to connect with others [12, 13] and enhance their social capital [14]. A fundamental process of the use of these sites is the posting of social and personal information by users [15]. Posted information typically includes personal data e.g. profile image, D.O.B., hometown and phone number [16]. While such information revelation is useful in connecting with others, there are issues of privacy control attached to the management of information, e.g. [17].

However, a potentially more hazardous consequence to users is the ability of information intended for one type of 'friend' leaking to other 'friends', or becoming publicly available. The dissemination of personal, sensitive or secret information by a social network 'friend' to their subsequent friend network (i.e. third-party to third-party sharing) greatly alters the balance of control and trust [8]. The release of personal information into the public realm can induce stress [18], and may cause problems for the subject of the disclosure with current relationships and job prospects. Moreover, managing the boundaries between the individual and different groups within their network can be complicated [17, 19], a problem known as 'conflicting social spheres' [17], or 'context collapse' [20]. Privacy controls are often difficult to find, implement or manage [19, 21].

We therefore contend that it is important to identify markers to information shared by users in tweets that is potentially sensitive to ensure an enhanced privacy

control mechanism and a greater user experience. E.g., identifying a status update as sensitive thus unsuitable for a particular person or group, could automatically trigger a user warning requesting consent to distribute. However, pop-up messages may be easily ignored. Alternatively, an algorithm may be used to automatically withhold sensitive messages from unintended recipients (pre-selected by the user). Although 71% of young users and 55% of older adults change their privacy settings in SNS from default [22], the use of a default option to categorize contacts and withhold sensitive content from certain groups may be preferable to an 'open' default.

1.3. Measurement of Self-Disclosure

Self-disclosure is difficult to measure, not only because of its complexity but because there is general disagreement over its definition and operationalization [11]. Methods of measurement include self-report surveys, behavioral observation, and objective metrics [11]. One of the few validated questionnaires, the Jourard Self-Disclosure Questionnaire (JSDQ) [23], consists of 60 items rendering it time consuming. The questions asked are about a participant's history of self-disclosure with a particular target [11], which limits its real-time application. Measurement using self-reports suffer from issues of impression management or social desirability.

An alternative method is to use content analysis, which can capture the depth of self-disclosure. Finer detail can be collected as each utterance can be assessed subject to its context, linguistic style and message content. Content analysis is influenced by the subjectivity of the researchers. For topics such as sensitive self-disclosure, an interpretative definition of 'sensitivity' is valid as sensitivity is subject to social norms. However, content analysis is time consuming, particularly with large samples. An aim of the present research is to identify sensitive self-disclosure using automated analysis, reducing the time required to identify concise areas within large samples. Content analysis can then be performed on the target areas, removing the need to manually filter entire texts, or markers used to compare corpora.

Linguistic analysis has been used to identify linguistic differences and markers in a variety of fields with complex applications including deception [24, 25] and emotion expression [26, 27]. The use of linguistic analysis to analyze the complex field of self-disclosure in NSM is a logical proposition.

To investigate our research aims we first use a comparison of tweets from Twitter and Secret Tweet in order to identify the linguistic markers for anonymous secrets. In study 2, we analyze tweets from the site 'Twitter' that have been rated according to sensitivity by six raters. We use Twitter as the core focus as it: controls for self-disclosure breadth by restricting the number of characters per post; is open and data are publicly available; it relies on social exchanges between users and their network of contacts, and tweets are considered to vary in the degree of self-disclosure.

2. Study 1 – Secret vs. Normal Tweets

2.1. Method

Two sites were chosen for the data collection in Study 1: Twitter (twitter.com) and Secret Tweet (secrettweet.com). Twitter is a micro-blogging site that allows users to 'tweet' a message up to 140 characters in length to the public twitter domain, and a list of contacts. Secret Tweet uses a similar interface and a 140-character limit, with the tagline "post your secrets to Twitter anonymously". The two sites were chosen specifically to control for message length, thus controlling for breadth of disclosure, measuring only differences in depth. Posts to 'SecretTweet' were treated as more sensitive than the typical post to Twitter. However, the sites do differ in a number of other ways – Twitter is not usually anonymous, while SecretTweet is, and Twitter users have specific 'followers' (although updates are usually publicly available), while posts to SecretTweet are listed on the web page of the site.

A comparison of tweets and secret tweets was conducted to form a between groups design with naturally occurring data. 'Normal' tweets, i.e. tweets from Twitter, were collected by copying & pasting results from a public search of "*a*" within the Twitter website. 250 tweets were copied at three periods across one day in May 2009 at 09:00, 14:00 & 18:00 to ensure users from across Western Europe and North America were collected at different periods of a weekday. Usernames and identifiable information were not processed. Care was taken to avoid ReTweets, celebrities, company profiles, marketing profiles and those written in languages other than English. These exclusion criteria were used to avoid people known publicly, re-use of others' posts, and to match tweets to secret tweets. Each tweet was from a unique Twitter Handle. Secret tweets did not need to be filtered as they were all posted anonymously on the main site page. 250 secret tweets were copied & pasted. A total of 500 tweets were collected (normal n=250, secret n=250). Demographic information including age, location, and gender are not available in Secret Tweet and were therefore not collected from Twitter.

The Linguistic Inquiry and Word Count (LIWC) [28] software is a widely used linguistic analysis tool. The LIWC is an advanced word counter that adds words in a given text to various linguistic categories. The percentage of words of a text input file is assigned to a particular category and given as an output variable. There are 80 categories pertaining to linguistic or psychological processes of language, or personal concerns [28].

The LIWC can only count words and word stems and therefore raters are used to verify if the two types of data differ in other qualities. Two naïve raters were used to assess the sensitivity and level of disclosure of each tweet to ensure that secret tweets were more sensitive than normal tweets. The raters were educated, native British English speakers, which could bias the results towards this cultural specific understanding of sensitivity. Tweets were rated on a scale of 1-10, where 10 referred to a high level of sensitivity or disclosure. The LIWC operator's manual was followed to clean the text [29]. Colloquialisms remained in the document to ensure future reliability of the results when used in real-applications. Care was taken to maintain the original structure and nature of the text. A Logistic Regression was performed to determine the predictive accuracy of the linguistic categories on tweet type. Significant predictors of tweet type are used as linguistic markers for self-disclosure of secrets.

Once the linguistic and statistical analyses had been performed, the tweets were examined to identify examples of the significant predictor word categories. If the analysis suggested either type of tweet contained more of a particular word category, the files were checked systematically to identify examples and determine their context, used in the interpretation of results and discussion.

2.2. Results & Discussion

The raters' scores showed a strong agreement for both sensitivity (α=.871, point-wise agreement) and level of disclosure (α =.836, point-wise agreement). The mean rater values and significant t-test indicate secret tweets to be more sensitive (secret=5.54, normal=1.98, t=-31.728, p<.000) and to have a higher level of disclosure (secret=6.56, normal=3.40, t=-27.504, p<.000) than normal tweets.

A logistic regression was performed with Tweet Type (normal or secret) as the DV and all 80 LIWC categories as predictor variables. A total of 500 cases (N=500) were analyzed (250 Normal Tweets, 250 Secret Tweets) and the full model significantly predicted Tweet Type (χ^2=500.11, df=16, p<.001). The model accounted for between 63.2% (at Step 1) and 84.3% of the variance in Tweet Type (at Step 16), with 91.6% of normal tweets predicted successfully, 92.0% of secret tweets predicted successfully and an overall prediction accuracy of 91.8%. A total of 16 predictors were derived from this analysis (Table 1).

Table 1. Linguistic markers of secrets

Word Category	Direction (β)*	Wald	Sig.
All Punct.	-0.053	6.964	0.008
Articles	-0.125	7.681	0.006
Exc. Marks	-0.629	19.541	0.000
Question Marks	-0.541	5.162	0.023
Fillers	-0.313	5.471	0.019
2nd Person (You)	-0.216	9.500	0.002
3rd Person Singular (She/He)	-0.227	10.479	0.001
Swear Words	-0.911	7.674	0.006
Personal Pronoun	0.298	52.272	0.000
Past Tense	0.083	4.501	0.034
Human Words	0.177	4.596	0.032
Inhibitions	0.262	6.111	0.013
Family Words	0.313	14.465	0.000
Word Count	0.100	9.106	0.003
Work Words	0.102	5.088	0.024
Sexual Words	0.300	9.913	0.002

**Positive value identifies Secret Tweets to the word category*

The β coefficients were examined to determine directionality (see Table 1). A positive β represents more of that word category for secret tweets. *Normal Tweets* contain more of the following word categories: *2nd Person (You), Articles, Swear Words, 3rd Person Singular (She/He)*, *Fillers*, *Question Marks*, *Exclamation Marks*, and *All Punctuation*. *Secret Tweets* contain more of the following word categories: *Word Count*, *Work*, *Personal Pronouns*, *Past Tense*, *Family*, *Human*, *Inhibitions*, and *Sexual*.

The strength and accuracy of the statistical model suggests that it may be possible to identify sensitive self-disclosure through the use of these 16 linguistic markers. These markers include personal pronouns, 3rd person singular words, family words, human words, and sexual words, and may relate to the typical content of secrets. Within the secret tweet data, disclosure examples discuss sensitive work issues, such as sexual

thoughts towards an employer. As such, this is represented by the increased use of sexual and work words in the observed results for secret tweets. Secret tweets showed a significant increase in the use of past tense, indicating that secrets relate to the previous events. Secrets often described an event that had happened rather than what the authors were doing presently, or planning to do, as in Twitter. E.g., a secret tweet states, "if my baby hadn't died, my husband would still love me." And a normal tweet focuses on the present, "… so tired right now. headed to campus, can't wait to come right back and sleep."

Secret tweets also contained significantly more words than normal tweets, suggesting more detail was involved, perhaps setting the scene, or adding elements to a story. This may be because the revelation of a secret on Secret Tweet is a planned activity, while normal tweets on Twitter may be more spontaneous. Previous research has shown support for differences in word count to be indicative of lying [24], or positive emotion [27]. However, word count alone does not indicate that the authors were showing increased positive emotion or deception. The 140-character limit of tweets may limit the sole use of this variable in identifying secretive self-disclosure.

Secrets contained less articles, swear words, question marks, exclamation marks, all punctuation, and filler words. This suggests a more formal sentence structure to secrets than normal tweets, due to less swearing and the succinct use of punctuation. This supports the idea that secret disclosure via SecretTweet may be planned and rehearsed. Furthermore, secrets may contain less swearing due to the confessional nature in which they are written. Identifying examples of secret and normal tweets in the data suggests that normal tweets tended to contain exuberant punctuation (e.g. multiple exclamation), a greater use of swear words, and greater use of 'leet'. An increased use of exclamation marks has been related to an increase in positive emotion within text [27]. This fits well with the current findings: secrets are less likely to contain positive emotion than normal tweets. Examples of secret tweets discuss being sorry, deaths, divorce, and adultery.

Secret tweets contained more personal pronouns than normal tweets. Interestingly, previous research has indicated that a decrease in self-oriented pronouns is related to deception and lying [24], suggesting that secrets are not only confessional nature, but may be the antithesis of deceptive communication.

Secret tweets can also be identified with fewer 3rd person singular (She/He) words and 2nd person (You) words. The use of references to others can be an attempt to dissociate the content of writing away from themselves [25]. As secret tweets in this data showed significantly less use of references to other people, it may indicate that the authors are involving themselves more often, relating the secret information towards their core constructs and private, sensitive informational attributes. This supports our earlier discussion and the rater agreement that secrets are more sensitive and thus useful to identify markers of sensitive self-disclosure.

The use of more family, human, inhibition, and sexual words was associated with secret tweets. These support the core categories of the social penetration model, typically involving aspects of sexuality, or family, or inhibitions [5]. Therefore, secrets relate to the core values of the self, non-secretive tweets relate less to core constructs.

Overall this method of analyzing and comparing the sensitivity and linguistic differences between two similar data sets has given a clear, constrained set of linguistic markers. However, we recognize that there are issues in comparing secrets posted openly and anonymously with tweets to a follower list by identifiable users. Secret tweets are bound to the context of the site and so conclusions about the linguistic markers obtained may be due to context rather than any difference in language use in comparison to normal tweets. The authors of tweets on each site may have different motives for disclosing.

3. Study 2 – Sensitive Self-Disclosure

Study 2 adopts the linguistic marker approach used in the first study, but addresses the problems inherent in comparing corpora from two different sites, albeit with the same character limit. Study 2 will address sensitivity within one site, Twitter, using ratings of sensitivity for the collected tweets. As Study 2 relies on the analysis of Tweet sensitivity within Twitter rather than comparison of two different corpora, more raters are used. Raters are asked only to rate the tweets for sensitivity, confusion was expressed when rating 'level of disclosure', in study 1. A final alteration is made to the methodology - in Study 2 we do not use all 80 LIWC categories but select a sub-group based on previous theory and literature. This reduces the likelihood of a Type I error.

3.1. Method

The same data collection method used in Study 1 for normal tweets was used in Study 2, and therefore the same collection and cleaning processes were employed. This resulted in 250 tweets used in Study 2.

Six raters were used to rate the tweets for their sensitivity. The raters were all educated, native British

English speakers of ages 23-42 years. Three were male, and three were female and each was familiar with the concept of Twitter. To reduce any bias from the researchers raters were not guided too rigorously in the definition of 'sensitive', but were instructed to "rate the following excerpts for how sensitive you consider them to be." Instructions were given as to the scoring of sensitivity, a scale of 1-10: 1 being 'not sensitive at all', and 10 being 'extremely sensitive'. A mean value of all six raters' scores of sensitivity were calculated and used in the analysis.

Selected variables from LIWC were chosen for the analysis, motivated by privacy, self-disclosure and information management literature. Specifically, we drew on the work on disclosure depth [5], language and intimacy [30], social spheres and context collapse [17, 20], the findings of study 1 and included general linguistic categories [28]. The following LIWC categories were used in a linear regression: all 30 linguistic process word categories [29]; *family*, *friend*, *humans*, *positive emotions*, *negative emotions*, *discrepancies*, *inhibitions*, *feel*, *body, health*, *sexual, work*, *home*, *money*, *religion*, and *death*. The 30 linguistic process categories were entered at Step 1 to identify markers of- and control for- any general linguistic variance. Theory-driven categories and those identified in Study 1 were entered at step 2.

3.2 Results & Discussion

Raters showed strong agreement for tweet sensitivity (α=.882, point-wise agreement). There was a main effect of word category on sensitivity of tweets at step 1 ($F_{(14,235)}$=3.229, p<.001) and step 2 ($F_{(30,219)}$=4.460, p<.001), with the model at step 1 accounting for 11.1% (R^2=.161, adjusted R^2=.111) of the variance observed and at step 2 accounting for 29.4% (R^2=.379, adjusted R^2=.294) of the variance observed. The variables in the model at step 2 demonstrate independent error values (Durbin-Watson = 1.850). Hereafter, the model is reported at step 2. Ten word categories predicted sensitivity to a confidence level of at least 95% (p<=.05, see Table 2).

Of these, five were linguistic processes (*3rd person singular, verbs, present tense, future tense* and *prepositions*) and five were theoretical predictors (1 personal concern: *death*; 4 psychological processes: *family, negative emotions, discrepancies* and *sexual*). One word category approached significance to a 95% confidence level, *3rd person plural* (p=.059, see Table 1). The tense word categories - *present* (β=-.352) and *future* (β=-.220) - had a negative relationship with sensitivity. The remaining 8 word categories had a positive relationship with sensitivity (see Table 2 for standardized βs).

Table 2. 10 Linguistic markers of sensitivity

Category Type	Word Category	+/-	β
Linguistic Processes	3rd person singular (She/he)*	+	.126
	3rd person plural (they)±	+	.104
	Verbs*	+	.406
	Present tense*	-	-.352
	Future tense**	-	-.220
	Prepositions	+	.171
Psychological Processes	Family**	+	.186
	Negative Emotions***	+	.243
	Discrepancies*	+	.126
	Sexual*	+	.129
Personal Concerns	Death***	+	.200

**p<.05. **p<.01. ***p<.001 ±p=.059*

The model accounts for 29.4% of the variance observed. This indicates that the model has reasonable strength, with ten linguistic markers accounting for the proportion of observed variance. Other external factors may contribute to this model that were not controlled for due to site design (e.g. age, gender, occupation and education level). The linguistic process categories that significantly predicted the tweet sensitivity were *3rd person singular (She/he), verbs* and *prepositions* (positive relationship), and *present* and *future* tense (negative relationship). *Third person plural (e.g. "they")* words approached significance (p=.059) with a positive relationship with sensitivity. Thus, more sensitive tweets had greater use of 3rd person plural words than less sensitive tweets, but the finding is only acceptable at a 94.1% confidence level. Further discussion will not include this variable. The significant variation in psychological processes and personal concerns had positive relationships with sensitivity, thus sensitive tweets contained more *family*, *negative emotion*, *discrepancies*, *sexual*, and *death* words.

The use of linguistic styles has been shown to relate to deception [24, 25], and emotional expression [26], amongst others. Therefore differences in general linguistic cues were expected in this similarly complex field of self-disclosure. The data set was examined systematically to identify tweets with high (>7/10) and low (<3/10) sensitivity to provide context to the linguistic markers. In the examples below marker

presence is italicized and the rating provided in brackets.

The increased use of 3rd person singular words (including she, he, him, her) is significantly related to more sensitive tweets. More sensitive tweets containing these word types include: "So *he* obsesses over me and *she* obsesses over *him* obsessing over me" (7/10) and, "Laugh out loud. and *he* just texted me and *he* said to tell you the truth, I've liked you for a while - i love my sister" (8/10). Compared to less sensitive tweets containing fewer he/she words, "anyone have a twitter app for windows mobile?" (2/10) and, "make it a great day" (2/10).

Verbs showed a positive relationship with sensitivity. The more sensitive tweets discussed more actions, "we *could* totally *be* friends again…. If I *could* talk to you face to face" (10/10), and, *"Having* a nervous break down! About to kill someone! I HATE MEN! UGH! I *can't wait* to *start* losing this weight so I *can* prove them all wrong!" (9/10). Less sensitive tweets contained fewer or no verbs, e.g., "a very good electro song" (3/10) and, "What a beautiful sunny day!" (2/10). The use of verbs in more sensitive tweets indicates that revealing one's **actions** is considered more sensitive. Less sensitive tweets did not disclose information that could identify a person's routine, location or intended actions. The exclusion of actions and the inclusion of the weather and musical opinion in the above examples of less sensitive tweets relate to the peripheral layers of the self-concept. This also supports the categories identified [5] within the context of Twitter.

Prepositions also showed a positive relationship with sensitivity with sensitive tweets containing words such as 'on', 'from' and 'beneath'. For example, "women *on* the train, *on* their way back home *from* work. Thinking thoughts. "am I a good mother?" (8/10), compared to the less sensitive tweet, "What a day… can it get any worse? Well, I hope not." (3/10).

Verbs and prepositions, although distinct, may be naturally correlated. Prepositions also relate to nouns and pronouns, which were not significant linguistic markers of sensitivity. However, personal pronouns significantly predicted secrets in study 1, suggesting they may be useful as general markers of sensitivity but differ relative to context. We don't suggest the presence of one maker be indicative of sensitivity in Twitter, but the markers should be used in combination generally.

Both *present* and *future* tense words were significantly negatively related to tweet sensitivity. This suggests that tweets in the present or future tense contain issues of low sensitivity. Secrets in study 1 were found to use more past tense. Although past tense was not significant in study 2, these findings suggest that higher sensitivity is either not related to a particular time frame, or sees a reduction of future and present words. In contrast, these findings could be due to the nature of Twitter, predisposed to releasing information about a user's current or intentional activities, e.g., "today will be a great day, I know!!!!" (1/10), and, "time to eat, and then tackle this mess of a backup solution" (2/10).

The theoretical predictors of sensitive self-disclosure (psychological processes and personal concerns) may represent deeper constructs than the use of syntax, author perspective or narrative voice (as indicated by differences in linguistic process words). The theoretical word categories, *family*, *negative emotion*, *discrepancies*, *sexual* and *death* predicted tweet sensitivity with a positive relationship. This is congruent with the social penetration model [5]. Core constructs relate to values and beliefs. E.g. family values can be observed in the closeness of kinship. In normal circumstances, family members are considered strong ties that maintain frequent and close communication.

The inclusion of family words in the tweets may also be due to the co-appearance of *death, sexual,* and *negative emotion* words. E.g., "is mourning with my dear relatives and friends tonight. The loss of a friend and brother in Christ is so hard to understand" (8/10), and, "Funny, how a weekend away from your spouse can be a great aphrodisiac. It's mandatory that couples keep some things separate" (6/10). The topic of death is sensitive, observed by the societal norms of western societies. Individuals are granted time off work for mourning, and funerals are a morbid and respectful send-off of the deceased. Death of a family member is more sensitive than the death of a stranger and may explain the co-appearance of family and death words, the examples and analyses above support this empirically.

4. General Discussion & Implications

Self-disclosure can lead to liking, reduced uncertainty and the development of relationships, and allows self expression [4-7]. However, over disclosure can lead to individuals being labeled as deviant and perceived as suspicious [4, 5], make recipients unsure as to the discloser's goals [7] and leave them feeling crowded [19]. The disclosure of sensitive personal information in public can leave individuals feeling stressed [18]. With the combined presence of the identified 10 linguistic markers in posts in NSM, it might be possible to identify over disclosure of sensitive information and reduce or prevent stress. E.g., young users of SNS may be forewarned of the potential

recipients of posts if they contain sensitive and potentially harmful self-disclosures. These harms can include further dissemination by others [8], over disclosure and the risk of being defriended [19], or information leakage across conflicting social spheres [17].

In NSM users are actively encouraged to share information [15]. Recipients can range from close friends and family to acquaintances and colleagues. Although 71% of users aged 18-29, and 55% of older users change their privacy settings from default [22], an automated, real-time process to select recipients of sensitive information is yet to be achieved. The linguistic markers found here provide the initial framework for this application.

This research suggests that markers of sensitive information relate to more core aspects of the self [5]. For secrets, the use of human, family, work and sexual words are considered more central as they relate to the core categories proposed by [5], and also led to deeper disclosure scores by our raters in study 2. The differences in linguistic markers for secrets and normal tweets could be due to the anonymity of Secret Tweet and the differences in site compared to Twitter. While the raters in study 1 agreed that secrets were more sensitive, they may have been written for a different purpose and audience compared to tweets from Twitter.

Study 2 identified linguistic markers for more and less sensitive tweets from within Twitter, and used only open profiles to prevent confounding results due to an account's openness. Therefore, differences in the markers of sensitive and non-sensitive tweets are considered more robust and less likely influenced by extraneous variables, e.g. anonymity and site design. Other extraneous variables may still be influential, e.g. age, gender, occupation and education level – none of which were available from the open twitter accounts. Using the number of followers, we explored analytically if the author's audience size predicted sensitivity. No effect was found.

With reference to our first aim, we have identified 10 linguistic markers of sensitive self-disclosure for use in future applications and research. By using a mean average of each marker within a given context, individual text excerpts can be compared to this baseline. A few examples of software application include the use of pop-ups, presenting an average sensitivity level of a user's messages through a GUI, allowing automated grouping of contacts as a default option in SNS, or to allow users to predefine who should receive more sensitive messages.

Our second aim was to contribute to the understanding of self-disclosure in a specific CMC context. The results of our research show that sensitive self-disclosure on Twitter relates to core aspects of the self rather than mundane aspects of social relations or the self. These core aspects related to sex, the family, discrepancies, negative emotions and death. Sexual and family words were significant predictors of sensitivity in both study 1 and 2, suggesting their application to sensitive self-disclosure across contexts.

The linguistic elements that contributed to self-disclosure research are the use of 3rd person singular (she/he), verbs, prepositions, future and past tense words. These markers demonstrate the differences in linguistic *style* by authors of more sensitive information. The presence of 3rd person singular words is interesting since we expected the key pronoun marker for sensitive disclosure to be personal pronouns (e.g. "I/me") – something we did find for secrets in Study 1, alongside the use of second person personal pronouns (you) in normal tweets. In this instance, the markers are related to the context of the tweets. This could be due to differences in the site and author visibility. It might be that sensitivity and disclosure in Twitter reflects its social nature – that is, sensitive disclosure is social in nature ("She's annoying") rather than self-related ("I'm annoyed"). Self-disclosure is usually performed to achieve a purpose [7]. The function of revealing secrets anonymously may be for self-clarification and personal expression. Whereas, more sensitive tweets (not secrets) may be disclosed for the purpose of social control, social validation or relationship development, thus involving third parties and deflecting away from the self. Disclosing information that relates to others might also be a way to disclose sensitive information without making the self too vulnerable, acting as a behavioral privacy control in light of the public nature of Twitter and limited knowledge of audience many users have.

The use of publicly available tweets in study 2 restricts these markers to sensitive self-disclosure within public CMC environments. The raters found that tweets within the public twitter domain can be sensitive. While the posting of personal information allows users to connect, posting sensitive information in public environments has been shown to induce stress [18]. This supports the need to identify sensitive text in real-time and help users control who is capable of viewing their messages. However, this assumes that users don't know their account is open – something we assume most Twitter users are aware of.

We also contribute to the ability to measure self-disclosure depth. Currently self-disclosure is measured through self-reports (typically history measures) or observable behavior [31]. The markers found in this research allow real-time identification of sensitive disclosure, and the ability to find examples of previous self-disclosure. A method to measure self-disclosure

depth is to adapt the 25-item JSDQ to produce scales of intimacy, e.g. [32], or use content analysis, like the raters in this research. These markers provide an alternative method for measuring depth of self-disclosure within Twitter.

The final aim of this paper was to provide the basis of a new tool to aid researchers of sensitive self-disclosure. The results of study 2 provide the basis for increasing the speed by which researchers can identify areas of interest within large text files. This will reduce the time spent analyzing the entire sample, and so researchers can focus their interest on the areas highlighted by the markers.

Overall we contend that automated linguistic analysis can add to the methods available to researchers identifying sensitive self-disclosure in online environments. This paper contributes to the literature on self-disclosure sensitivity in a specific CMC context and provides the basis for the development of automated processes, reminders, and contact grouping.

4.1. SensiTweet

The linguistic markers identified in Study 2 were used to develop a proof of concept application to indicate to users the level of sensitivity of their tweets (for a given Twitter handle) compared to a baseline average. This application presents one example of how the markers could be used to aid users via a GUI. It has not been verified or tested at this stage.

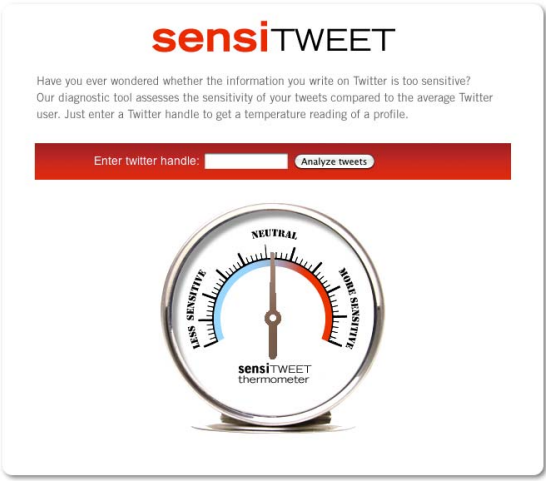

Figure 1. SensiTweet GUI.

A total of 84,000 tweets were scraped from public accounts of Twitter and stored in a database. An average score for each marker across the 84,000 tweets was calculated. A php script is used to process up to 2000 Tweets on demand. Feedback is given to users via the interface shown in Figure 1. The application can be found at http://interactionslab.net/sensitweet.

4.2. Limitations & Future Research

We recognize several limitations with this research and the potential for future work to be conducted. (1) The use of only open Twitter accounts may further restrict the context of the tweets collected and a comparison of markers for open and closed accounts may be useful. (2) The markers are currently valid for use within Twitter. Comparisons of markers between Twitter and other CMC may be useful to further validate the markers, or a longitudinal study may be useful to capture greater variance within author's writing styles.. (3) We recognize that the use of just 2 raters in study one limits its findings, however this was addressed for study 2 with the use of 6 gender balanced raters. (4) The use of these markers to highlight instances of sensitive disclosure in large texts currently relies on the use of LIWC, a future application to highlight text within files may be preferable and faster. (5) SensiTweet is merely a proof of concept, we have not validated the markers against the Tweets it captures. It does, however, illustrate our point regarding the ability to capture the markers in real-time.

5. References

[1] Worthy, M., A.L. Gary, and G.M. Kahn, *Self-disclosure as an exchange process.* Journal of Personality and Social Psychology, 1969. **13**(1): p. 59-63.

[2] Joinson, A.N. and C.B. Paine, *Self-disclosure, privacy and the Internet*, in *The Oxford Handbook of Internet Psychology*, A.N. Joinson, et al., Editors. 2007, Oxford University Press: Oxford, UK. p. 237-52.

[3] Smyth, J.M., *Written emotional expression: Effect sizes, outcome types, and moderating variables.* Journal of consulting and clinical psychology, 1998. **66**(1): p. 174-184.

[4] Berger, C.R. and J.J. Bradac, *Language and Social Knowledge. Uncertainty in Interpersonal Relations*. The Social Psychology of Language 2, ed. H. Giles1982, London, UK: Edward Arnold Ltd.

[5] Altman, I. and D.A. Taylor, *Social penetration: The development of interpersonal relationships*1973, USA: Holt, Rinehart and Winston, Inc.

[6] Cozby, P.C., *Self-disclosure: A literature review.* Psychological bulletin, 1973. **79**(2): p. 73-91.

[7] Derlega, V.J. and J. Grzelak, *Appropriateness of Self-Disclosure*, in *Self-Disclosure: Origins, Patterns, and*

Implications of Openness in Interpersonal Relationships, G.J. Chelune, Editor 1979, Jossey-Bass: San Francisco, Washington, & London. p. 151-176.

[8] Houghton, D.J. and A.N. Joinson, *Privacy, Social Network Sites, and Social Relations.* Journal of Technology in Human Services, 2010. **28**(1): p. 74-94.

[9] Afifi, T.D., J. Caughlin, and W.A. Afifi, *Exploring the dark side (and light side) of avoidance and secrets.*, in *The dark side if interpersonal relationships*, B. Spitzberg and B. Cupach, Editors. 2007, Erlbaum: Mahwah, NJ. p. 61-92.

[10] Spiekermann, S., J. Grossklags, and B. Berendt, *E-privacy in 2nd Generation E-Commerce: Privacy Preferences versus actual Behavior*, in *ACM Conference on Electronic Commerce*2001: Tampa, FL, USA. p. 38-47.

[11] Chelune, G.J., *Self-disclosure: Origins, patterns, and implications of openness in interpersonal relationships*1979, San Francisco, Washington, & London: Jossey-Bass.

[12] Chen, G.M., *Tweet this: A uses and gratifications perspective on how active Twitter use gratifies a need to connect with others.* Computers in Human Behavior, 2011. **27**(2): p. 755-762.

[13] Joinson, A.N., *'Looking at', 'looking up' or 'keeping up with' people? Motives and uses of Facebook*, in *CHI 2008 - Online Social Networks*2008: Florence, Italy. p. 1027-1036.

[14] Ellison, N., C. Steinfeld, and C. Lampe, *The benefits of Facebook" Friends:" Social Capital and College Students' Use of Online Social Network Sites.* Journal of Computer-Mediated Communication, 2007. **12**(3): p. 1143-1168.

[15] Burke, M., C. Marlow, and T. Lento, *Feed Me: Motivating Newcomer Contribution in Social Network Sites*, in *CHI 2009 Conference*2009: Boston, MA, USA. p. 945-954.

[16] Gross, R. and A. Acquisti, *Information revelation and privacy in online social networks*, in *Proceedings of the 2005 ACM workshop on Privacy in the electronic society*2005, ACM: Alexandria, VA, USA. p. 71-80.

[17] Binder, J., A. Howes, and A. Sutcliffe, *The Problem of Conflicting Social Spheres: Effects of Network Structure on Experienced Tension in Social Network Sites*, in *CHI 2009*2009: Boston, MA, USA. p. 965-974.

[18] Little, L. and P. Briggs, *Private whispers/public eyes: Is receiving highly personal information in a public place stressful?* Interacting with Computers, 2009: p. 316-322.

[19] Joinson, A.N., et al., *Digital Crowding: Privacy, Self-Disclosure and Technology*, in *Privacy Online. Perspectives on Privacy and Self-Disclosure in the Social Web*, S. Trepte and L. Reinecke, Editors. 2011, Springer: Heidelberg and New York. p. 31-44.

[20] Marwick, A.E. and D.M. Boyd, *I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience.* New Media & Society, 2011. **13**(1): p. 114-133.

[21] Bonneau, J. and S. Preibusch, *The Privacy Jungle: On the Market for Data Protection in Social Networks*, in *Workshop on the Economics of Information Security*2009: University College London.

[22] Madden, M. and A. Smith, *Reputation Management and Social Media: How people monitor their identity and search for others online*, 2010, Pew Internet & American Life Project. Accessed at http://pewinternet.org/Reports/2010/Reputation-Management.aspx.

[23] Jourard, S.M. and P. Lasakow, *Some factors in self-disclosure.* Journal of Abnormal Psychology, 1958. **56**(1): p. 91-98.

[24] Hancock, J., et al., *On Lying and Being Lied To: A Linguistic Analysis of Deception in Computer-Mediated Communication.* Discourse Processes, 2008. **45**(1): p. 1-23.

[25] Newman, M.L., et al., *Lying Words: Predicting Deception From Linguistic Styles.* Personality & Social Psychology Bulletin, 2003. **29**(5): p. 665-675.

[26] Gill, A.J., et al., *The Language of Emotion in Short Blog Texts*, in *Computer Supported Cooperative Work 2008*2008. p. 299-302.

[27] Hancock, J., C. Landrigan, and C. Silver, *Expressing emotion in text-based communication*, in *Proceedings of ACM Conference on Human Factors in Computing Systems (CHI 2007)*2007. p. 929-932.

[28] Pennebaker, J.W., M.E. Francis, and R.J. Booth, *Linguistic Inquiry and Word Count (LIWC): LIWC2001.*2001, Mahwah, NJ: Lawrence Erlbaum Associates.

[29] Pennebaker, J.W., R.J. Booth, and M.E. Francis, *Operator's Manual, Linguistic Inquiry and Word Count: LIWC2007*2007, Austin, Texas: LIWC Inc.

[30] Gilbert, E. and K. Karahalios, *Predicting Tie Strength With Social Media*, in *CHI 2009*2009: Boston, MA, USA. p. 211-220.

[31] Chelune, G.J., *Measuring Openness in Interpersonal Communication*, in *Self-Disclosure: Origins, Patterns, and Implications of Openness in Interpersonal Relationships*, G.J. Chelune, Editor 1979, Jossey-Bass: San Francisco, Washington, & London. p. 1-27.

[32] Altman, I. and W. Haythorn, *Interpersonal Exchange in Isolation.* Sociometry, 1965. **28**(4): p. 411-426.